DTIC FILE COPY

(1)

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

AD-A197 385

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Naval Ocean Systems Center 6b. OFFICE SYMBOL (if applicable) NOSC	7a. NAME OF MONITORING ORGANIZATION Naval Ocean Systems Center
6c. ADDRESS (City, State and ZIP Code) San Diego, CA 92152-5000	7b. ADDRESS (City, State and ZIP Code) San Diego, CA 92152-5000
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Office of Naval Research 8b. OFFICE SYMBOL (if applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State and ZIP Code) Independent Research Programs (IR) OCNR-10P Arlington, VA 22217-5000	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO.: 61152N PROJECT NO.: ZT87 TASK NO.: AGENCY ACCESSION NO.: DN308 055

11. TITLE (include Security Classification)

EXPLORING THE BACK - PROPAGATION NETWORK FOR SPEECH APPLICATIONS

12. PERSONAL AUTHOR(S)

S. Luse, D. Martin, S. Nunn, J. Waters

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Presentation/speech	FROM Apr 1988 TO Apr 1988	June 1988	

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	neural networks back-propagation

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Neural networks have sophisticated abilities for processing and filtering signals. In particular, Elman and Zipser demonstrated that the back-propagation network develops significant feature representations which may be useful for both segmenting and recognizing speech. Such networks might find applications in speech compression and/or normalization. The network's apparent potential for speech applications justifies further exploration, and this paper describes our work in progress.

Presented at Speech Tech '88, 25 Apr 1988, New York, N.Y.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED
22a. NAME OF RESPONSIBLE INDIVIDUAL S. Luse	22b. TELEPHONE (include Area Code) (619) 553-3656 22c. OFFICE SYMBOL Code 441

DD FORM 1473, 84 JAN

83 APR EDITION MAY BE USED UNTIL EXHAUSTED
ALL OTHER EDITIONS ARE OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

EXPLORING THE BACK-PROPAGATION NETWORK FOR SPEECH APPLICATIONS

Stephen Luse
Doug Martin
Stephen Nunn
Jeff Waters

Human Factors and Speech Technology Branch
Code 441
Naval Ocean Systems Center
271 Catalina Blvd.
San Diego, Ca. 92152-5000

ABSTRACT

Neural networks have sophisticated abilities for processing and filtering signals. In particular, Elman and Zipser [1] demonstrated that the back-propagation network develops significant feature representations which may be useful for both segmenting and recognizing speech. Such networks might find applications in speech compression and/or speech normalization. The network's apparent potential for speech applications justifies further exploration, and this paper describes our work in progress.

1. Background

Fundamental problems remain unsolved in speech processing. Although recent developments have occurred in digital signal processing techniques and their hardware implementations, many questions are unanswered concerning correct perceptual units of speech, the use of prosodic information in the speech signal, and viable voice compression algorithms.

Neural networks provide a powerful means for pursuing answers to these questions. While performing complex mappings, these networks extract fundamental information about the input signal. Relying on this ability, networks may be trained to aid in the accomplishment of difficult speech processing tasks.

One particular network paradigm, the back-propagation network, has the demonstrated capability to represent significant features of speech [1] [3]. The back-propagation network is a three-layer neural network that behaves as an interpolative-associative memory. It has the ability to learn mappings by example; that is, it will generalize input-output pair relationships. The mappings that are learned are dependent on the input-output pairs that are used during the network's training period [2].

The network's capability may be useful for isolating and modifying a variety of speech features, and might find immediate application in speech compression (to improve low data rate voice transmission), speech normalization (to improve voice recognition), and speech adaptation (for voice feature modification). In short, the back-propagation network is a powerful tool for analyzing and manipulating speech.

2. Purpose

The goal of our research is to explore how back-propagation networks, trained to learn the significant representations of preprocessed speech, affect novel speech data. Using networks of different sizes with different preprocessing methods, we hope to discover the features learned and how this information may aid the performance of difficult speech processing tasks, such as:

1) Identifying phonemes in the context of continuous speech for both speaker dependent and independent recognition;

2) Capturing speech characteristics of a particular speaker for use in speaker recognition and verification;

3) Compressing and encoding the speech signal for low data rate transmission as well as secure communications;

4) Capturing critical vocal tract parameters and speaker prosodic information such as pitch, stress, rate, etc., to be used for disguising a person's speech or changing the identity of the speaker.

We also hope to gain insight into the amount of processing power required for useful applications. Limited processing power restricts the size of our networks and increases the amount of time necessary to train each network.

3. Network Design

The size of the back-propagation network must be designed so the network can learn to represent the fundamental aspects of the preprocessed speech. Of the three layers in our network, the input and output layers are the same size (they have the same number of processing elements), so that we can perform a one-to-one mapping of speech data.

The hidden layer is less than one-half the size of the input or output layer. This standard "hour glass" design forces the network to learn significant representations of speech. The network must determine these significant features in order to compress the data in the hidden layer and recreate it with reasonable accuracy.

Other researchers [1] have illustrated that a network, with a hidden layer size of less than one-half the size of the input layer, is capable of learning significant representations of speech. Our research will explore the use of networks with hidden layers of this size and smaller.

4. Preprocessing Methods

A network may have difficulty drawing meaningful features from the variable speech signal, since there are practical limitations to the complexity of the mappings that a neural network can accomplish. Much research in speech recognition and perception has involved the attempt to find a representation for the speech that would simplify the problem of the great variability in the speech signal [4]. Since each representation emphasizes a different aspect of the speech signal, each will affect in a different manner a network's performance, including training time, error rate, and feature development.

Existing theories of speech recognition suggest at least three useful representations of the speech signal [5]. These representations are described below along with the preprocessing methods which can partially capture features important to the representations:

A) *Speech Perception Theory -- Preprocessing: Raw Speech*

Recognition can be achieved by extracting speech features, such as voice onset times and formant transitions, that have been experimentally established as being important to the human perception of speech. These time domain features are present in raw digitized speech.

The advantages of using raw speech for training the network are 1) it is easy to acquire, 2) it imposes no preprocessing perturbation of the signal, and 3) raw speech must contain all vocal tract parameters and frequency domain information in some form. No information has been eliminated. One possible disadvantage is that, since raw speech does no pre-encoding, it may place excessive burdens upon the network during training.

B) *Speech Reception Theory -- Preprocessing: FFT*

The human auditory process can also be modeled by extracting parameters and classifying patterns as done in the ear, auditory nerves, and sensory feature detectors. These parameters and patterns are found principally in the frequency content of the signal.

Frequency-domain representation of speech information is doubly advantageous. First, acoustic analysis of the vocal mechanism shows the production of critical formant frequencies that permit a concise description of speech sounds [6]. Second, a great deal of evidence indicates that the ear makes a crude frequency analysis at an early stage in its processing [6].

Furthermore, FFT data is relatively easy to extract from the time-domain signal, and there is a great deal of information in the FFT signal. However, there also is a great deal of irrelevant information in an FFT.

C) *Speech Production Theory -- Preprocessing: LPC*

Speech can be recognized by understanding the method of speech production--the parameters describing the vocal tract. Important parameters include vocal tract resonances, rate of vibration of the vocal cords, and manner and place of articulation.

Linear Predictive coding analysis is a powerful technique for estimating the basic speech parameters, such as pitch, formants, and vocal tract area functions. LPC is based on the idea that a speech sample can be approximated with a linear combination of past speech samples. By minimizing the differences between the actual speech samples and the linearly predicted ones, a unique set of predictor coefficients can be determined.

The advantage of using LPC is that it provides an extremely accurate estimate of the speech parameters. The main disadvantage is that the data in each segment of speech has been enormously reduced to a representation that contains only a few filter coefficients.

5. Approach

For our experiments, we train the network to perform a one-to-one mapping, using speech from one male speaker. Once the network is trained to an acceptable error level, we process speech from three other male speakers as well as from the training speaker. (For purposes of limiting training time and obtaining preliminary results, an acceptable level was set at 10% error. In general, an acceptable error is the lowest error possible, perhaps 1% or 2%, with a reasonable amount of training time, probably less than 72 hours.)

To determine the effect of the network, we use a speech processing system whose performance can be measured with or without the presence of the neural network (see Figure 1). First, the system error is measured *without* the network. Digitized speech is preprocessed. Inverse preprocessing is performed and error measurements are made by comparing the input and output speech. Second, the system error is measured *with* the network. A back-propagation network is trained to perform a one-to-one mapping of the preprocessed speech. The network output is inverse preprocessed to obtain an approximation of the input speech. Error measurements again are made by comparing the input and output speech. Together, these measurements allow us to determine the network's contribution to the system error.

With a one-to-one mapping, a mean-square error measurement can be used to compare input and output speech waveforms. Suppose the input waveform is x_n and the output waveform is y_n, both of length N samples. The error signal e_n, is defined as:

$$e_n = y_n - x_n$$

The energy in the error signal E_e, is defined as:

$$E_e = \frac{1}{N}\sum_n (\bar{e} - e_n)^2$$

where $\bar{e}$ is the average error over the N samples. It is useful to express the error energy as a percentage relative to the energy in the input signal S:

$$E = \frac{E_e}{S} = \frac{\frac{1}{N}\sum_n (\bar{e} - e_n)^2}{\frac{1}{N}\sum_n (\bar{x} - x_n)^2}$$

where $\bar{x}$ is the average input signal. While this error measure is not as useful as a correlation function, it is easy to calculate and provides necessary insight into the network's behavior.

This error measure can also be used to compare the output speech to the training speech. If the network has decreased the overall error between these two signals, it appears the network has learned some of the unique features of the training speech and it is imposing those features on the input signal.

For further study, we may listen to the effect of the network on novel speech and, based on the results, train new networks to emphasize such effects. We hope to design several demonstrations of potential applications to illustrate the network's capabilities.

6. Initial Studies

We began our studies by training a network to map raw speech, using the short word, "zero". We wanted some initial measure of training time and intelligibility. After training a 50-20-50 network (fifty processing elements in the input layer, 20 processing elements in the hidden layer, and fifty processing elements in the output layer), we found training time was a matter of hours and that the processed speech was readily intelligible.

After this initial test, we experimented with network training using batch processing--the network interconnection weights were adjusted only after all of the input patterns in the word "zero" were processed, rather than adjusting the weights after each pattern. Batch processing provided a more reliable training error measurement, although training was slower.

Using batch processing, we trained networks of various sizes on the word "zero". Rough training error measurements are shown in Table 1.

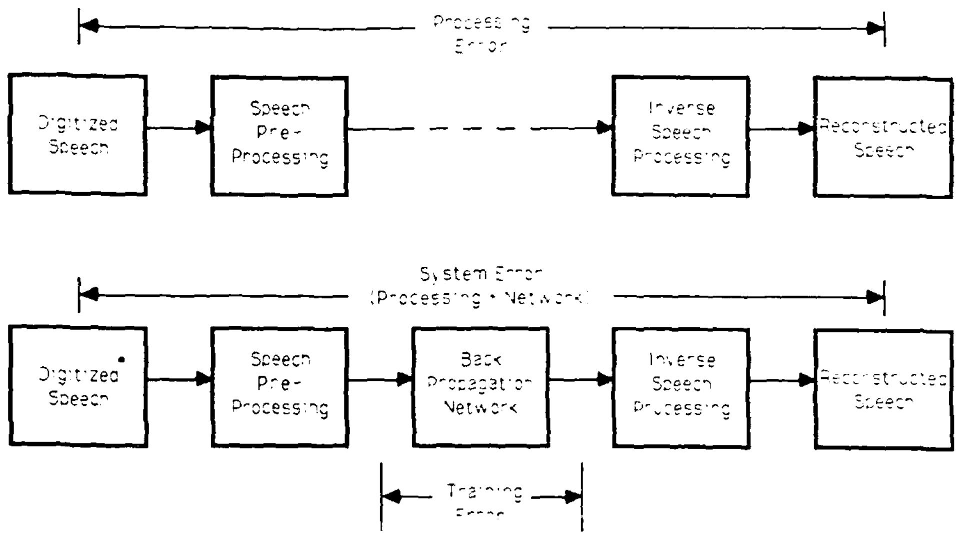

Figure 1. System Used for Error Measurement

In general, the larger the network, the lower the error. It's difficult to say whether the larger networks performed better because they received more significantly-sized chunks of speech data, or simply because they were larger and had more connections in which to store more information.

Our initial training studies verified the conclusion reached by other researchers that the back-propagation learning rate follows a power-law relation. The first graph in Figure 2 shows pass error v. time for seven networks of different sizes, while the second graph shows regression lines for the same data. The correlation coefficient for all regression lines is 0.96 or greater, indicating a strong fit to the power-law.

As another initial experiment, we trained a network on one person's voice and then input a different person's voice through the network. What would the output sound like--the trainer's voice or the input voice? The output sounded like the input voice, not the trainer's voice. This was interesting, since it suggested the network was learning general features of speech, not unique features of a particular voice. Yet, our current results, described below, suggest the network may be trying to learn unique features as the network's hidden layer size decreases.

Table 1. Initial Studies

Training Error vs. Time		
Network Size	Error 10 hrs.	Error 20 hrs.
64 - 20 - 64	14%	10.5%*
128 - 20 - 128	14	11.5
40 - 30 - 40	17	9.7*
64 - 40 - 64	13	8.0
128 - 40 - 128	8	3.5
128 - 60 - 128	8	4.23
192 - 60 - 192	7	2.3*
*Estimated		

As a result of initial studies, we decided that a network size of 64-20-64 would be an appropriate starting point for our experiments. We also learned that the length of utterances had a severe impact on the required training time. When longer training inputs were tested, we determined that non-batch processing was more efficient for our purposes (training in hours as opposed to batch training in days). We therefore chose non-batch processing with the utterance "zero" as our test word and recorded the utterance from four male speakers with an 8 kHz sample rate. The length of the digitized waveforms varied from 4,864 samples to 5,888 samples, representing 0.61 seconds to 0.74 seconds in duration.

For raw speech, the input was segmented into 64-point blocks, each representing 16 ms of the original signal. Each block of 64-point samples was presented to the input layer.

For FFT data, the input was segmented into 128-point blocks, each representing 32 ms of the speech. After performing the FFT, each resulting 64-point block, representing the magnitude of the spectrum, was presented to the input layer. (Phase information provided by the FFT is currently unused; however, further experiments using phase information are planned.)

7. Current Results

Using one speaker's utterance as a training set, we trained four back-propagation networks to within 10% training error. Table 2 shows the resulting training errors for each network.

Table 2. Training Error

Training Errors for Four Networks		
Network Size	64-20-64	64-10-64
Raw Speech	9.8%	9.4%
FFT Speech	9.9%	10.0%

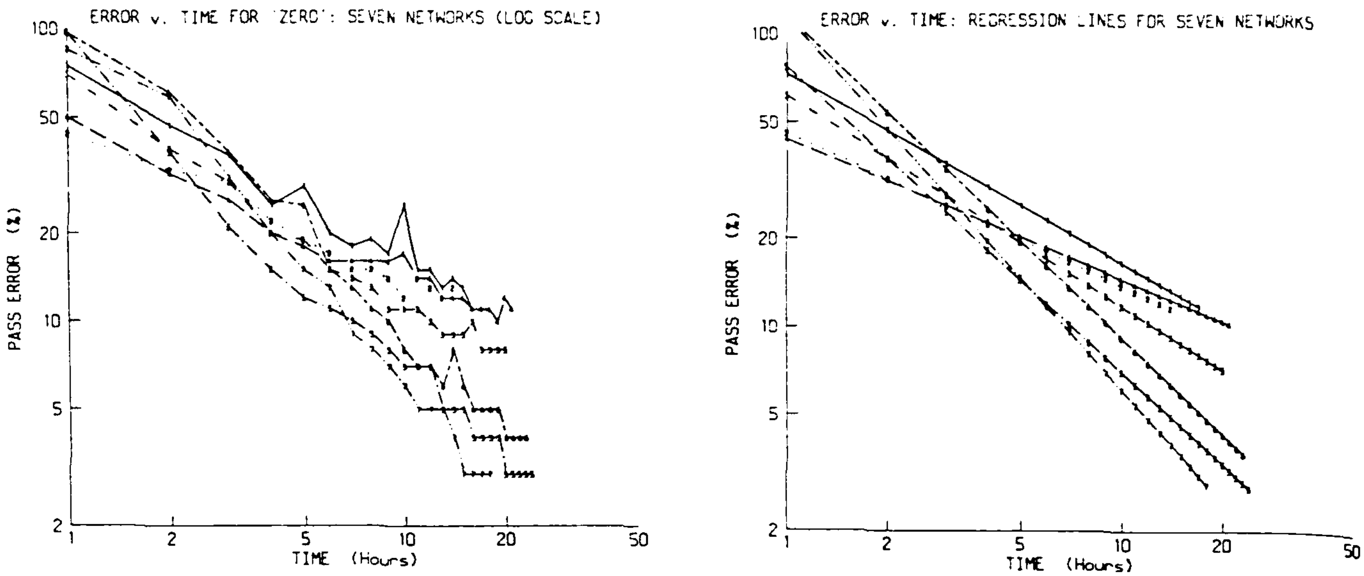

Figure 2. Training Error vs. Time (for seven networks)

After training the networks, we processed the training speaker data and the data for the other three speakers. The resulting errors are shown in Table 3 for all four networks along with the processing error for comparison.

In order to determine if the network learned features relative to the training data, we made error measurements between the test speaker's output speech and the training speaker's original speech. The results are shown in Table 4.

8. Observations

Although it is too early to make firm conclusions, some specific observations concerning the initial and current results are provided below:

-- *Training Rate:* Networks learn the most in the first few hours of training, and only learn gradually thereafter. (See Figure 2.)

-- *FFT Error:* FFT preprocessing (and inverse preprocessing) alone instills a 15-25% error in the signal, in addition to any error instilled by the network. (See Table 3, the column entitled "Processing Error".)

-- *Mapping input speech to input speech:* The back-propagation system using raw speech mapped with a lower error (26-33%) than did the system using FFT data (37-50% error). If one subtracts processing error from the system error, then the network trained on FFT data contributes roughly the same amount of error as the raw speech case. (See Table 3.)

-- *Mapping input speech to training speech:* We wanted to see if the network decreases the error between a test speaker's voice and the training speaker's voice (making, in some sense, the test speaker's voice more similar to the trained speaker's voice). Table 4 shows this measurement for the networks. The system using FFT data mapped with a lower error (123-130%) than did the back-propagation system using raw speech (170-189% error).

-- *Varying hidden layer size & mapping input speech to input speech:* Using raw speech, the system mapping degraded (the error increased by approximately 2.5%) as the hidden layer size dropped from 20 to 10 processing elements. Using FFT data, the system mapping improved (the error decreased by approximately 3%) as the hidden layer size dropped from 20 to 10. (See Table 3.)

-- *Varying hidden layer size & mapping input speech to training speech:* Using raw speech, the system mapping improved (the error decreased by approximately 3%) as the hidden layer dropped from 20 to 10 processing elements. Using FFT data, the system mapping degraded (the error increased by approximately 6%) as the hidden layer size dropped from 20 to 10. (See Table 4.)

-- *Compression:* As shown in Table 3, both systems (raw and FFT) do an acceptable mapping of input speech to input speech, with a data compression ratio as high as 6.4 to 1.0. The mapping is considered acceptable when the error is below 70%, since our subjective listening suggests that the speech is intelligible with this amount of error.

-- *Mapping many voices to one:* As shown in Table 4, the system using FFT data, when compared with the system using raw speech, did a better mapping of all voices to the training voice. Although the error instilled by the system is immense (over 100%), reduction in the error is expected in future tests as the networks are allowed to train below 10% error.

-- *Learning the speaker's voice:* The FFT system appears to do the better job of learning significant representations of speech which are unique to the training speaker's voice. The system error in Table 4 suggests that the input voices were modified slightly to resemble the speaker's voice. Our subjective listening suggests this may be the case, although application of phase during reconstruction of the signal appears to have an equally significant effect on voice identification features. If the FFT system error can be reduced, the FFT system may serve a useful role in voice modification.

-- *Learning and the hidden layer size:* Different sizes of hidden layers appear to perform different functions for voice adaptation, depending on the preprocessing method. For raw speech, the additional compression of the data as the hidden layer size shrinks does slightly aid the attempt to model the trainer's voice. For FFT networks, the additional compression actually hinders the attempt, increasing the error range. Apparently, additional compression for raw speech aids the learning of significant features unique to the training speaker, but hinders this function with the FFT data.

9. Current Conclusions

From the above observations, we conclude the following: (1) Whether a network is trained on FFT or raw speech data makes no difference in the error contributed by the network. (2) The back-propagation network can compress speech effectively. In our case, 6.4-to-1. Further compression may be possible with larger networks. (3) The network has difficulty learning significant representations of speech *which are unique to the speaker's voice*. Our data suggests, however, that a network trained on FFT power-spectra does have potential in this regard. Note that phase information, which contains information relevant to the identity of a speaker, was not used during training to keep the size of the networks manageable and commensurate with the processing power we have available.

Our results, although informative, would be more useful if we could devise a speech processing system that allows the network to learn features with a more acceptable training error (1 or 2%). Our experience shows this could be possible if: (1) more processing power were available to decrease training time, and (2) larger networks could be used. Both of these issues can be addressed by modifying the gradient descent algorithms used in training the networks, as well as running our experiments on faster machines. Both possibilities are being considered.

10. Future Study

The remainder of our research for this year will be to apply LPC speech data to the network. The LPC data sets will be much smaller and may provide the network with a feature set that allows for better feature extraction. The results of this work will be available by late summer of this year.

Table 3. Comparing Output with Input Speech

Error Between *Input* and Output Speech		Processing Error	System Error (Processing + Network)	
Network Size			64-20-64	64-10-64
Raw	Training Speaker	0.002%	9.8%	9.5%
	Speaker 2	0.002	28.1	30.5
	Speaker 3	0.003	31.0	33.6
	Speaker 4	0.002	26.2	29.2
FFT	Training Speaker	19.9%	32.9%	30.0%
	Speaker 2	21.5	48.9	45.1
	Speaker 3	15.5	40.7	37.9
	Speaker 4	24.7	49.7	45.4

Table 4. Comparing Output with Training Speech

Error Between *Training* and Output Speech		Without Network	With Network	
Network Size			64-20-64	64-10-64
Raw	Training Speaker	0.0%	9.8%	9.5%
	Speaker 2	210.	189.	186.
	Speaker 3	190.	174.	167.
	Speaker 4	196.	173.	170.
FFT	Training Speaker	0.0%	32.9%	30.0%
	Speaker 2	210.	123.	130.
	Speaker 3	190.	124.	129.
	Speaker 4	196.	124.	130.

11. References

[1] Elman, J. L., & Zipser, D. (1987). Learning the Hidden Structure of Speech. University of California at San Diego, Institute for Cognitive Science.

[2] Luse, Stephen A. "Neural Networks for Speech Applications." Speech Technology Magazine, Oct./Nov. 1987.

[3] Sejnowski, T. J., & Rosenberg, C.R. "NET-talk: A parallel network that learns to read aloud," Johns Hopkins University Department of Electrical Engineering and Computer Science Technical Report 86/01 (1986).

[4] Klatt, Dennis H. "The Problem of Variability in Speech Recognition and In Models of Speech Perception," in **Invariance And Variability In Speech Processes**. ed. Joseph S. Perkell & Dennis H. Klatt (1986).

[5] Lea, Wayne A. **Trends In Speech Recognition**. Prentice-Hall (1980).

[6] Flanagan, J.L. **Speech Analysis, Synthesis, And Perception**, 2nd ed. (1983)

LMED

8